Understanding Domain Registration Abuses

Scott E. Coull[a,*], Andrew M. White[b], Ting-Fang Yen[c], Fabian Monrose[b], Michael K. Reiter[b]

[a] *RedJack, LLC. Silver Spring, MD 20901*
[b] *University of North Carolina, Chapel Hill, NC 27599*
[c] *RSA Laboratories, Cambridge, MA 02140*

Abstract

The ability to monetize domain names through resale or serving ad content has contributed to the rise of questionable practices in acquiring them, including domain-name speculation, tasting, and front running. In this paper, we perform one of the first comprehensive studies of these domain registration practices. In order to characterize the prevalence of domain-name speculation, we derive rules describing "hot" topics from popular Google search queries and apply these rules to a dataset containing all .com registrations for an eight-month period in 2008. We also study the extent of domain tasting throughout this time period and analyze the efficacy of ICANN policies intended to limit tasting activity. Finally, we automatically generate high-quality domain names related to current events in order to measure domain front running by registrars. The results of our experiments shed light on the methods and motivations behind these domain registration practices and, in some cases, underscore the difficulty in definitively measuring these questionable behaviors.

Keywords: domain names, speculation, front running, tasting

1. Introduction

Domain names have become an integral component of modern web browsing by allowing users to navigate to web sites using memorable phrases and keywords. In fact, many users will often assume that domain names based on intuitive keywords will direct them to the desired web site, known as *type-in navigation*. For this reason, domain names have become quite valuable, which has led to a variety of practices where domain names are opportunistically registered simply to profit from them. One such dubious domain registration practice is *domain speculation*, where a domain name is registered with the intention of reselling it for a profit at a later date or generating ad revenue from type-in navigation traffic [1]. Though speculation is technically allowed by Internet Corporation for Assigned Names and Numbers (ICANN) rules, it has led to more abusive behaviors by registrars and speculators. For instance, *domain tasting* allows a speculator to register large numbers of domains at no cost for a short grace period during which she can assess the potential value of the domain. Another example is *domain front running*, where domain registrars use queries about domain availability made by their users to preemptively register domains then subsequently resell them to those same users for a profit. The security problem underlying these behaviors is not unlike that presented by spam during its emergence, in that both activities take advantage of loopholes in existing policy to profit from unintended uses of the respective systems. While the security and legal communities have identified certain behaviors as clear abuses of the registration process (*e.g.,* typosquatting [2, 3, 4, 5, 6]), the practices and impact of domain speculation, tasting, and front running are still not well-understood.

In this paper, we perform the first in-depth study of questionable domain name registration activity, including a characterization of domain speculation, an analysis of the prevalence of domain tasting, an investigation of the possibility of domain front running by popular domain registrars, and an analysis of the impact of ICANN policies on these abusive behaviors. Specifically, we used popular search terms provided by Google to develop association rules that describe keywords that are likely to occur together in domain names related to current events and "hot" topics. These association rules were then used to generate regular expressions that searched for domain speculation activity, and to automatically generate domain names used to measure domain front-running activities. Through our analysis of all .com registrations during an eight-month period in 2008, we shed light on the prevalence of these abusive registration practice, their motivations, and the inherent difficulties in detecting them.

*Corresponding Author
Email addresses: scott.coull@redjack.com (Scott E. Coull), amw@cs.unc.edu (Andrew M. White), tingfang.yen@rsa.com (Ting-Fang Yen), fabian@cs.unc.edu (Fabian Monrose), reiter@cs.unc.edu (Michael K. Reiter)

2. Related Work

The inherent importance of the Domain Name Service (DNS) in enabling navigation of the web makes it a natural target for attackers seeking to misdirect users to malicious web sites. Due to their prevalence and potential impact on users, these misdirection attacks have been widely studied. For example, a handful of recent studies focused on measuring the prevalence of typosquatting [2, 3, 4, 5, 6] and homograph [7, 8] attacks, which take advantage of the user's inability to differentiate the intended domain name from what appears on the screen.

Unfortunately, studies of the less malicious, yet still questionable, domain registration activities that we examine in this paper appear to be limited primarily to a series of status reports by ICANN. Their analysis of domain tasting [9] examined the use of the five-day add grace period between June 2008 and August 2009. Overall, they observed a significant decrease in tasting activity after a temporary provision was instituted in July 2008 that limited registrars to a relatively small number of no-cost registration deletions. Similarly, their preliminary statement on domain front running activities [10], based on user complaints, found that the majority of claims were due to user error or oversights during the registration process. A follow up ICANN report [11], incorporating a study of 100 randomly generated domains, found no evidence of front-running activity by registrars. The obvious limitations of that investigation are its relatively small scale and the use of randomly generated names that were easy to identify and ignore.

More recently, the LegitScript organization performed a study of fraudulent online pharmaceutical web sites and their connections to particular domain registrars [12]. Their results show definitive connections between certain registrars and prevalent domain registration abuses. Meanwhile, a study by Liu *et al.* [13] on the impact of registrar-level intervention mechanisms shows that these methods fail to stem the tide of domain registration abuses. Finally, Moore *et al.* [14] study the use (and abuse) of popular Google search terms to opportunistically drive traffic to content-free web pages containing only ads and malware. The results of their study indicate that it is possible for popular search engines to limit this activity by removing so-called "low-quality" web pages from their rankings, which opens up a potential avenue for mitigation of the domain registration abuses discussed in this paper.

3. Preliminaries

To successfully achieve our goals we needed to overcome two challenges. The first lies in decomposing Google search queries about a given topic into combinations of keywords that are likely to appear in domain names related to that topic. Broadly speaking, we assume that the searches that users make on Google about an event or topic are closely related to the domain names they would navigate to using type-in navigation. These type-in navigation domains are prime targets of domain squatters and front runners, and therefore the focus of our investigation. The second challenge lies in developing a method for determining which domains are pertinent to our study. Before proceeding further, we describe the data sources and methods used to address these challenges.

Data Sources. For our study, we made use of a variety of data sources, including both historical domain name registration data and longitudinal information on Google search query popularity. Our historical analysis of domain name registrations was based on data made available through VeriSign's Zone Access Program [15] which contains 62,605,314 distinct .com domain registration events from March 7, 2008 to October 31, 2008. The VeriSign data contains domain names, their associated name servers, and the date of each registration event. The data also contains information about de-registration events, which we used in our analysis of domain tasting behaviors. For the remainder of the paper, we refer to the set of all domain name registrations contained in the VeriSign data set as the *background* set.

Furthermore, in order to gain a sense of the popularity of various topics or events, we made use of data provided by Google via its Insights for Search and Trends services. These services rank the top searches made by users over a given time frame, and provide up to ten related searches for each. Our methods assume that these queries adequately represent the hot topics that caused their increase in popularity in the Google search engine, and that this increase in search popularity is an indicator of the desirability of domains related to the hot topic. In our study, data from the Insights for Search service was used to derive rules for searching domains in our VeriSign data, while the Trends service provides real-time search rankings that were used to generate high-quality domains names for our domain front-running experiment. A *topic* in our study is defined to be a top ranked search, along with its ten related searches. Each search is composed of a set of keywords.

Unfortunately, due to the sensitive nature of the domain registration data, it was not possible for us to obtain more recent data since our initial study [16]. Based on recent reports provided by VeriSign [17] and anecdotal evidence of the problems surrounding .XXX and generic top-level domain (gTLD) registrations [18, 19], we believe that many of the issues identified within the 2008 dataset persist today. In particular, recent studies by Liu *et al.* [13] and Moore *et al.* [14] have identified similar levels of abusive domain speculation activities in more recent data.

Association Rule Mining. Even though the Google searches reflect trending news and events, they often contain unrelated keywords, as well. Performing a direct string matching of those search terms on our background set would likely return many irrelevant domains. Hence, to identify combinations of keywords that best represent the topic

associated with them, we applied association rule mining techniques [20]. These techniques consider an itemset $I = \{i_1, \ldots, i_m\}$ containing all items that can be found in a transaction, and a transaction set $T = \{t_1, \ldots, t_n\}$, where t_α is a set containing the items associated with the αth transaction. The *support* of a set of items X is defined as $supp(X) = n_X/n$, where n_X is the number of transactions containing all items in X. An *implication* between sets of items X and Y, denoted as $X \Rightarrow Y$, indicates that the presence of the items in X implies items in Y will also be present. The *confidence* of an implication $X \Rightarrow Y$ is defined as $conf(X \Rightarrow Y) = supp(X \cup Y)/supp(X)$. An implication is considered to be a *rule* if the sets have a sufficient level of support and confidence.

For our purposes, we used the notions of support and confidence defined above to decompose each Google query into groups of keywords specific to the topic at hand. To do so, we considered each search for a given topic to be a transaction with each keyword in the search acting as an item in the transaction's set. We then decomposed those searches into sets of co-occurring keywords based on the confidence of the keywords' pairwise implications. Specifically, we first examined each ordered pair of keywords in the search query to discover all of the bidirectional rules (*i.e.,* implications where the confidence in both directions was above our threshold), and merged them by assuming transitivity among the implications. These bidirectional rules describe the groups of keywords that must appear together in order to be meaningful to the given topic.

Next, we augmented the rule set by examining unidirectional implications, which indicate that the antecedent of the implication should only be present where the consequent also exists. As before, we assumed transitivity among the rules to merge them appropriately. If a keyword was not the antecedent in any rules, we added it as a singleton set. The algorithm returns the union of the rule sets for each of the search queries, which contain all of the groups of keywords that represent the topic associated with those searches. A more detailed description of the rule mining algorithm can be found in Algorithm 1.

Due to the inherently noisy nature of the data used in our study, it is important to carefully set thresholds used in our rule mining and other selection procedures. The threshold selection methodology is complicated by the fact that our data provides no notion of what values might be related to a given topic (*i.e.,* the data is unlabeled). Therefore, we used cluster analysis techniques to automatically set the thresholds used in our study, rather than appealing to manually derived thresholds. Specifically, we made the observation that we needed to separate only two classes of unlabeled values: those that are interesting with respect to our analysis and those that are not. Thus, to determine a threshold we first used the k-means++ algorithm [21] with $k = 2$ to partition the unlabeled values into the sets S_1 and S_2 (*i.e.,* interesting and uninteresting), and set the threshold as the midpoint between these two clusters.

4. Domain Name Speculation

Our first objective is to examine the relationship between new domain registrations and so-called "hot" topics in an effort to gain a better understanding of domain speculation. To do so, we followed an iterative process that consists of: (i) generating rules that are specific to the topic at hand, (ii) converting those rules into regular expression to select domains, and (iii) pruning and verifying the set of selected domains to ensure they are, in fact, related to the topic.

First, we gathered Google Insights data for each month in 2008, and treated the set of searches related to the topics as transactions, which were used in our association rule mining algorithm to generate rules. Recall that a threshold confidence value dictates which implications in our set of transactions should be considered rules. To determine this threshold, we calculated the confidence between all pairs of keywords within a topic and used the threshold selection method discussed in Section 3 on these values to set the appropriate threshold. The resulting set of rules were further pruned to ensure that non-specific rules were discarded. To do so, we scored each rule r_i for a topic as $S(r_i) = \sum_{k \in r_i} supp(k) \times |k|$, where r_i is a rule for the current topic (represented as a set of keywords), $supp(k)$ is the support of keyword k, and $|k|$ denotes the string length of the keyword k. Intuitively, this procedure produces rules for a topic that contain predominately long, important keywords, and removes those rules that may introduce irrelevant domains due to more general or shorter keywords. Again, we used the threshold selection method to set a threshold score for the rules associated with a topic, where all rules above that threshold were retained.

Given the high-quality rules generated for each topic, we converted them to regular expressions by requiring all keywords in a rule to be found as a substring of the domain, and that keywords in bidirectional implications appeared in their original ordering. To add a level of flexibility to our regular expressions, we also allowed any number of characters to be inserted between keywords. The domains selected by the regular expressions for a given topic undergo one more round of pruning wherein the domain was assigned a score equal to the sum of the scores for each of the rules that matched it. These domain scores were given as input to the threshold selection algorithm, and any domains with scores above the threshold were manually verified to be related to the associated topic. These related domains are herein referred to as the *relevant* set. Table 1 shows an example of the conversion process from search query, to rule sets, and finally to regular expressions used to search our domain registration data.

Results. Our search methodology selected 21,103 distinct domain names related to 116 of the 120 hot topics from the VeriSign dataset. Of these, 15,954 domains in 113 topics were verified to be directly related to the topic at hand (*i.e.,* the relevant set). The percentage of relevant

Algorithm 1 Search Query Mining (S, t)

Input: S – set of searches; t – confidence threshold
Output: *rules* – set of association rules for the searches S

```
Initialize rule set for current topic, rules
for all searches s ∈ S do
    Initialize rule set for current search, R

    // First, we find the transitive closure of bidirectional implications
    for all pairs of keywords (k1, k2) ∈ s do
        if conf({k1} ⇒ {k2}) ≥ t and conf({k2} ⇒ {k1}) ≥ t then
            if a rule r containing k1 or k2 already exists in R then
                Extend the rule r to include the new keyword
            else
                Add the rule k1 ⇔ k2 to R
            end if
        end if
    end for

    // Next, expand rule set R with unidirectional implications
    for all pairs of keywords (k1, k2) ∈ s do
        if conf({k1} ⇒ {k2}) ≥ t and conf({k2} ⇒ {k1}) < t then
            if a rule r containing k1 already exists in R then
                Add a new rule to R that contains all elements of r as well as k1 ⇒ k2
            else
                Add a new rule k1 ⇒ k2 to R
            end if
        end if
    end for

    // Finally, add the keywords not associated with implications as singleton rules
    for all keywords k ∈ s do
        if a rule r containing k does not exist in R then
            Add a new singleton rule k to R
        end if
    end for

    Add R to rules
end for
return rules
```

domains per topic, averaged over all topics, is 91%. Overall, these results indicate that the majority of our rules selected high-quality domain names; a small number of topics produced very general rules, often because of unrelated or non-specific Google search queries.

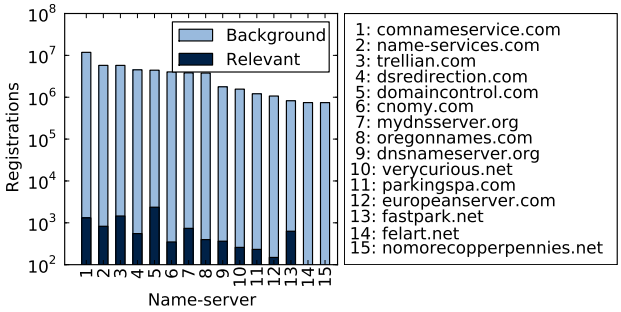
In order to discover the unique properties of the potentially speculated domains that our methodology selected, we examined several features and compared them to those of the background set of domains. First, we looked at the distribution of registrations among the name servers and registrars within the background and relevant sets, respectively. In Figure 1(a), we show a log-scale plot comparing the background and relevant domain registrations associated with the top fifteen name servers in the background set. For clarity, we also provide the distribution of the top name servers in the relevant domain set in Figure 1(b). Clearly, the distribution of registrations over these two sets is significantly different as evidenced by the ranking of name servers and the comparison plot in Figure 1(a). In fact, when we took a closer look at the name servers, we found that the majority of those in the relevant set are associated with domain parking [22] services, whereas the background set contains a much smaller fraction.

Similarly, we compared the top registrars from the background distribution to those from the relevant set, as shown in Figure 2. To characterize the distribution of registrars for background domains, we used the VeriSign monthly reports for the .com top-level domain (TLD) to derive the number of domains registered by the top fifteen registrars over the eight-month period examined in our study. Analyses reveal that some registrars, such as GoDaddy and eNom, maintain their popularity as registrars in both sets. Upon closer inspection, we also found significant differences. For example, Network Solutions drops precipitously from a rank of three to ten in the relevant set, and several registrars are exclusively in the relevant set. These findings indicate that some registrars are clearly preferred by speculators as the name servers above were, albeit to a lesser extent.

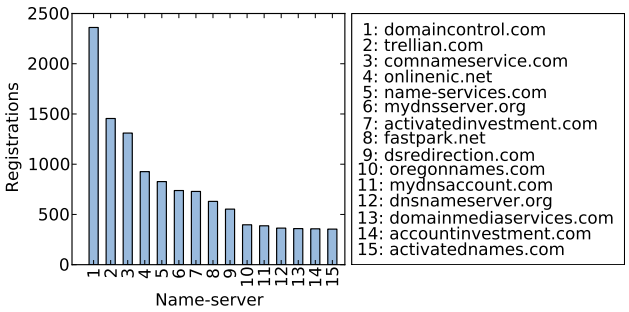
Timely Registrations. A potentially interesting subset of our relevant domains are those that were registered soon after the popularity of the associated hot topic or current event in Google's search queries. While an intuitive notion of timeliness is to simply examine how soon after the initial spike in Google search popularity a domain was registered, certain topics that have periodic or ongoing levels

Google Query	Association Rule	Regular Expression
phillies world series	phillies ⇒ (world ⇔ series)	(world.*series) & (.*phillies.*) or (world.*series)
clinton vs obama	vs ⇒ clinton ⇒ obama	(.*obama.*) or (.*clinton.*) & (.*obama.*) or (.*clinton.*) & (.*obama.*) & (.*vs.*)

Table 1: Decomposition of a Google query into rules and regular expressions.



(a) Name servers in the background set.



(b) Name servers in the relevant set.

Figure 1: Comparison of name servers between background and relevant domain sets.

of popularity do not easily fit this model, such as the topic "Obama", which is shown in Figure 3.

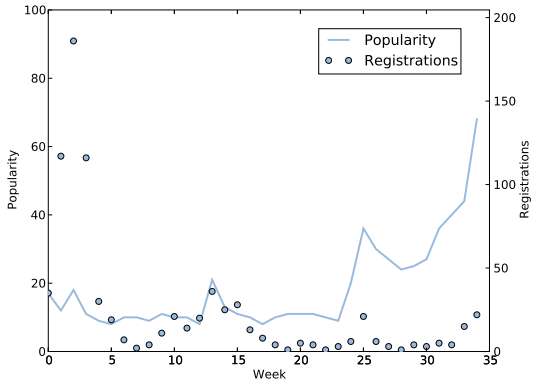


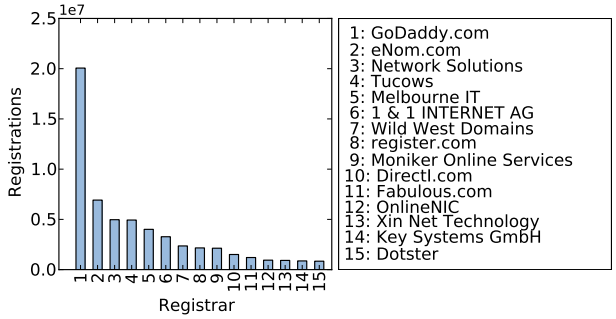
Figure 3: Popularity and registration time-series for topic "Obama."

A more appropriate and general measure of timeliness is the *cross-correlation lag* between the time series of relevant registrations for a topic and the popularity of that topic's Google queries. Generally speaking, cross-correlation lag is a measure of how many time steps one sequence must be delayed in order to maximize the cross-correlation between the two series. This amounts to maximizing the dot-product between the two series by sliding the second series forward and backward in time. Formally, the cross-correlation between two discrete time series $x(t)$ and $y(t)$ at delay d is defined as $\phi(d) = \sum_{t=0}^{n} x(t) * y(t-d)$, where n is the length of the two series, t is the current time step for series $x(\cdot)$, and d is the lag. Then, the cross-correlation lag is defined as $\arg\max_d \phi(d)$ [23]. A negative lag value means that the registration series precedes the popularity series, while a positive lag indicates that the registration series succeeds the popularity series.

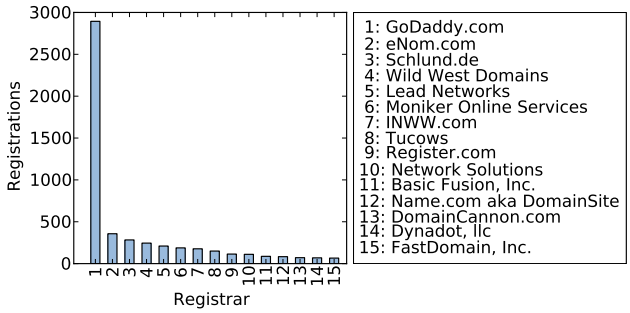
Given this notion of timeliness, we examined all domains with a cross-correlation lag value between -1 and +1 weeks. The set of domains associated with topics that fall within this range is denoted as the *timely* set. This timely set contains 7,574 domain names associated with 52 distinct topics. For each of these domain names, we performed the same set of measurements as our previous analysis of relevant domains to tease out any trends that may exist among timely domains. In Figure 4, we compare the distribution of registrars and name servers between those domains in our relevant and timely sets. While it is obvious that there are certain registrars and name servers that are more likely to produce timely domains related to hot topics, the overall distribution of these two sets is essentially the same.

5. Domain Tasting

The second form of questionable domain registration behavior that we examine is domain tasting, where a registrar is allowed to delete a domain at no cost within five days of the initial registration, also known as the add grace period. This policy can be easily abused by registrars and registrants alike in order to gain information about the value of a domain via traffic statistics taken during the grace period. To study the prevalence of domain tasting, we selected all domain names from the background set of domains that were registered and then deleted within five days; we refer to these as the *tasting* set.

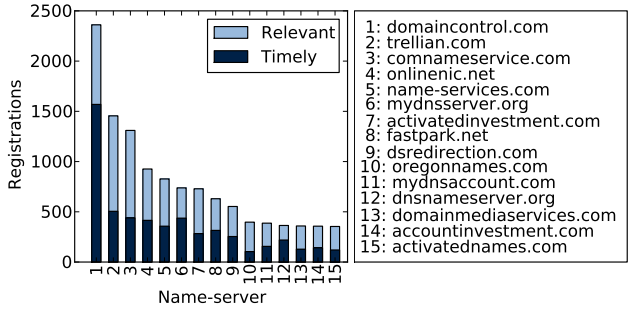


(a) Registrars in the background set.

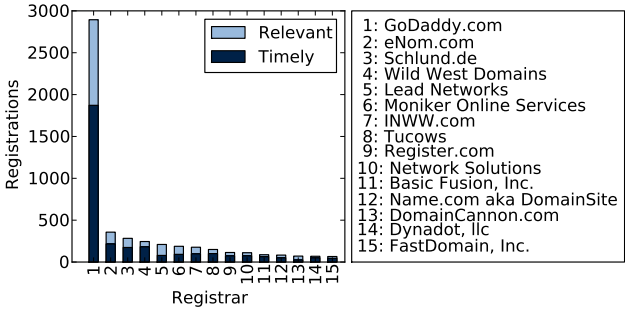


(b) Registrars in the relevant set.

Figure 2: Comparison between registrars for background and relevant domain sets.



(a) Registrations for top name servers in the relevant and timely set.



(b) Registrations for top registrars in the relevant and timely set.

Figure 4: Comparison of name servers and registrars between relevant and timely domain sets.

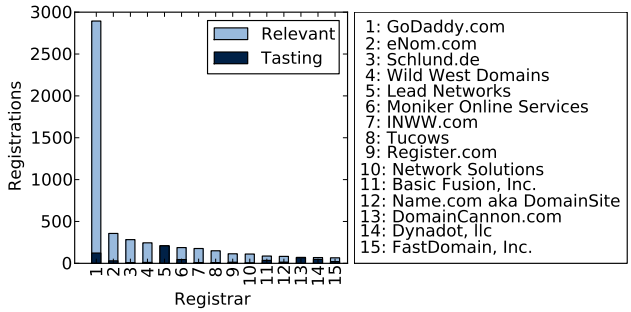
Results. From the full VeriSign dataset (*i.e.,* background set), we identified 47,763,141 (76%) distinct registrations as the result of domain tasting, with 10,576 of those occurring in our relevant set of potentially speculated domains (66% of the relevant set). On average, these tasting domains were registered for 3.4 days before being deleted under the no-cost grace period policy. Figures 5(a) and 5(b) show the comparison of registrars and name servers between all relevant domains and those relevant domains involved in tasting activity. The graphs clearly illustrate that these relevant tasting domains are strongly connected with particular registrars or name servers, in some cases representing all of their registrations.

In June 2008, ICANN made changes to their policies in order to limit the practice of domain tasting. Specifically, the new policy charges companies for excessive domain de-registrations above a certain threshold, making domain speculation expensive [9]. These changes took effect on July 1st, 2008, which positions us perfectly to provide an independent analysis of the impact of this policy change on the tasting of .com domains. For our purposes, we split the background dataset into a *pre-reform* period and a *post-reform* period. We found 42,467,156 pre-reform tasting registrations with an average duration of 3.4 days, and 6,290,177 post-reform registrations with an average duration of 3.8 days. For our relevant domains, there is a similar proportion of tasting registrations with 9,270 pre-reform registrations and 1,433 post-reform registrations. These relevant tasting domains were registered for an average of 2.8 and 3.7 days, respectively. In both the background and relevant tasting domains, there is a clear trend toward longer registration periods after the enactment of tasting reform.

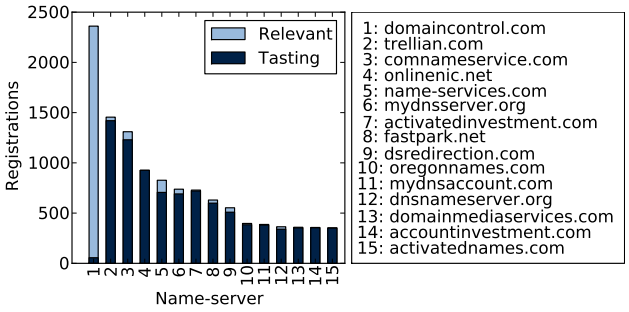
To examine the impact of the reform on the top fifty registrars and name servers in the background and relevant domain sets, we examined their change in rank after implementation of the new tasting policies. Figure 6 shows the change for names servers and registrars associated with the relevant set. Notice the substantial drops in rank for those name servers and registrars occupying the middle ranks (*i.e.,* positions 10-40 in the pre-reform data). Although several of the top-ranked name servers in both the background and relevant sets are predominantly associated with tasting domains, they are able to maintain – or even improve – their rank despite the drop in tasting registrations (*e.g.,* trellian.com).

6. Domain Front Running

Finally, we explore the extent of domain front-running activities among the top domain registrars. To do this successfully, we needed to generate relevant (and presumably

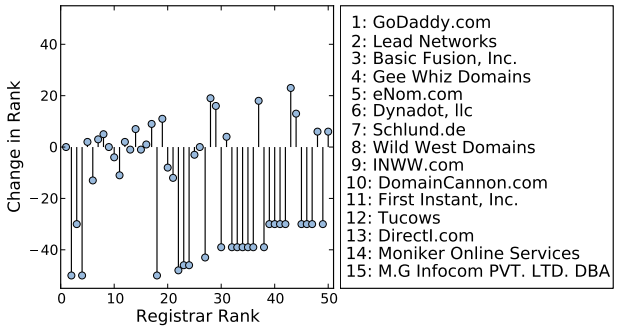


(a) Registrars in the relevant set.

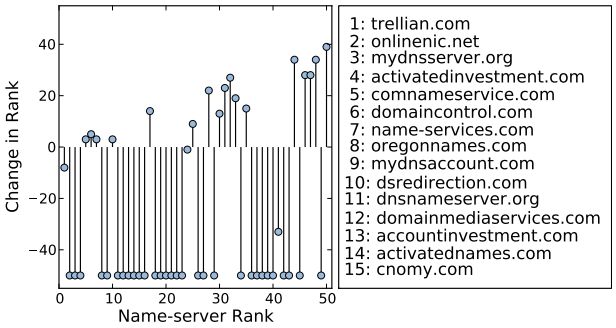


(b) Name servers in the relevant set.

Figure 5: Comparison of name servers and registrars for tasting in the relevant set.



(a) Registrar rank in the relevant set.



(b) Name server rank in the relevant set.

Figure 6: Changes in pre-reform rank for name servers and registrars in the relevant domain set.

desirable) domain names for very timely topics, then query domain registrars for the availability of those domains in a manner that simulated widespread interest.

Our approach for generating domain names is similar to that of the rule generation procedure. We began by gathering search queries from the top two popularity classifications (*i.e.,* "volcanic" and "on fire") for the current day from Google Trends, and used those searches as transactions in our rule mining process. As before, we set confidence and pruning thresholds for the rule generation for each topic separately using the threshold selection procedure described in Section 3. At the end of this process was a set of rules for each hot topic for the day.

For each association rule, we created domain names containing the keywords in the bidirectional implications of the rule in the order in which they appear in their original Google search. We then augmented this domain name to generate additional names by creating all permutations of it with the keywords in the unidirectional implications. For singleton rules, we used the keyword by itself as the domain name string. Additional domains were generated by appending popular suffixes to the initial domains (*e.g.,* "blog," "online"). Table 2 provides a concrete example of the domain names generated by our methodology.

The generated domains were divided among the registrars in our study such that no two registrars received the same domain name. The domains for each registrar were further divided into queried and held-out sets. This division of domains allowed us to examine the increase in the rate of registration for those domains that were sent to registrars over those that were not, and pinpoint the increase in registration rate for certain domains to a particular registrar. Furthermore, in order to ensure that our queries appeared to emanate from a diverse set of locations, we made use of the PlanetLab infrastructure. We distributed domains for each topic to between two and four randomly selected nodes, which then queried the registrars for availability of these domains via the registrars' web sites. Lastly, each day we checked `Whois` records to determine if any of our queried domains were subsequently registered. In this experiment, we assumed that a statistically significant increase in registration rate between queried and held-back domains by a particular registrar is related to front-running activities.

Results. In our study, we issued queries as described above to the nineteen most popular registrars, accounting for over 80% of the market share, according to RegistrarStats.com. Over the period spanning December 1^{st} 2009 to February 1^{st} 2010, we generated 73,149 unique domains of which

Google Query	Association Rule	Domain Name
phillies world series	phillies ⇒ (world ⇔ series)	worldseries.com, phillieswordseries.com, worldseriesblog.com
clinton vs obama	vs ⇒ clinton ⇒ obama	obama.com, clintonobama.com, obamavsclinton.com

Table 2: Decomposition of a Google query into rules and domains.

60,264 (82%) were available at the time of generation. Of those available at the time of generation, 16,635 were selected for querying and distributed to the PlanetLab nodes, leaving 43,629 domains in the held-back set. A total of 23 of the queried and 50 of the held-back domains were registered during this period.

To examine the significance of our results, we performed statistical hypothesis tests for each of the registrars in isolation. Specifically, we modeled the rate of registration in both the queried and held-back case as a binomial distribution with probability of success equal to the unknown rate of registration. The Fisher-Irwin exact test was applied instead of the standard z-test, since it avoids approximation by a normal distribution and explicitly calculates the probabilities for the two binomials given the numbers of queried and held-out domains. Our analysis indicates that none of the registrars are associated with a statistically significant ($p < 0.05$) increase in the registration rate of queried domain names. Possible reasons for the lack of evidence in front running behavior are discussed next.

7. Summary

In what follows, we discuss the implications from our empirical analyses and examine the relationship between tasting, front-running, and speculation activities.

On the Quality of the Generated Rules. Based on the results of our speculation and front-running experiments, we argue that the rule mining and threshold selection methodologies worked surprisingly well given such noisy data. For our speculation experiments, we found that an average of 91% of the selected domains were related to their respective popular topics, and many of our automatically generated domains were indeed registered. For those rules that generated non-relevant domains, the primary cause can be attributed to incoherence in the related search terms provided by Google. Nonetheless, we believe that our techniques show significant promise in taking unstructured keywords and returning general rules that can be applied to a variety of problems.

It is interesting to note that we found anecdotal evidence of some domain registrants using similar techniques to register domains, particularly while the ICANN domain tasting policy was still lax. For one, we found strong temporal correlation among "hot" events, their rise in Google search rankings, and the volume of domain registrations related to those events. Moreover, when we took a closer look at the data, we found obvious patterns of automated domain registration associated with certain name servers. As an example, we found one registrant associated with name servers from verycurious.net, whose general approach was to generate domain names by selecting a disease (for example, Abdallat Davis Farrage Syndrome) or city name, and then append relevant phrases to them (*e.g.,* `abdallat-davisfarragesyndromrecovery.com`). These domains were often registered in batches of tens of thousands at a time, and then de-registered five days later. Even with stricter tasting policies in place, such automated methods can still be used, albeit with a more restricted set of domain names.

Incentives for Misbehavior. A natural question that arises when considering these abusive domain registration behaviors is: what are the incentives that drive them? To begin to answer this question, we performed a cursory analysis of the contents of potentially speculated domains selected by our methodology, along with an examination of potential ad and resale revenue associated with the topics in our study. Most of the domains that we examined contained significant pay-per-click ad content, and our analysis shows that many of these sites were hosted by known domain parking firms. Based on data gathered from Google's AdWords Traffic Estimator, we found that the average cost-per-click for the topics in our study was $0.76 per click, and many of these topics have expected click rates in the 300-400 clicks per day range. Beyond ad revenue, domain names associated with the topics we studied were resold for an average price of $1,832, according to domain-name auctions from DomainTools.com, with the largest of these being $15,500 for `obama.net`.

Clearly, there is significant financial incentive to both resell popular domains and to use parking services to generate advertising revenue. In fact, as long as the average revenue among the domains owned by the speculator exceeds the hosting and registrations costs, the speculator is better off retaining as many domains as possible and only serving ad content. As a concrete example, we note that the keywords associated with the automatically generated domains from our front-running study would have produced revenue in excess of $400 per day (again, based on Google's Traffic Estimator), while domain parking services can be purchased for as little as $3.99 per domain each month. This represents a net profit of approximately $11,700 per month from ad revenue alone for the 73 registered domains in our front-running study. Furthermore,

Keyword	Cost/Click		Clicks/Day		Cost/Day	
	low	high	low	high	low	high
obama	$0.90	$1.25	207	259	$186.30	$323.75
mccain	$1.26	$1.89	256	256	$322.56	$483.84
tiger woods	$0.49	$0.61	207	259	$101.43	$157.99
haiti	$0.64	$0.80	291	364	$186.24	$291.20
toyota	$1.50	$1.87	6,142	7,677	$9,213	$14,355.99

Table 3: Ad revenue estimates for selected keywords as of June 2009.

Keyword	Domain	Date Sold	Marketplace	Price
obama	obama.net	10/24/2009	WickedFire	$15,500.00
	obamacenter.com	6/15/2008	Sedo	$251.00
	obamaforpresident2012	2/23/2009	SnapNames	$119.00
mccain	mccain-lieberman.com	2/12/2008	Sedo	$250.00
beijing 2008	chinaolympics2008.net	3/21/2008	NameJet	$82.00
	chinaolympics2008.org	3/21/2008	NameJet	$77.00

Table 4: Domain re-sale history for select domains as of December 2009.

the strong connection between domain popularity and revenue provides insights into the use of tasting and front-running behaviors as a mechanism for determining the true market value for domains without having to invest capital. Tables 3 and 4 give some examples of cost-per-click ad revenues and domain re-sale values for topics in our experiments.

Difficulty of Measurement. Another surprising lesson from our study is that many of these questionable registration behaviors are particularly difficult to definitively measure. In the case of speculation, we attempted to use several metrics to distinguish those domains registered due to speculation from those registered for legitimate use, including the length of registration, the timeliness of the registration after the increase in search popularity, the rate at which hosting changes, and manual inspection of web page content. Of these, only inspection of the content yielded any significant results, and even in this case there were several instances where it was difficult to identify the true purpose of the page (*i.e.,* to deliver legitimate content, or to serve ads). Our experiences in differentiating legitimate from abusive content mirror those of Moore *et al.* [14], who appeal to several content-based features to uncover abusive web pages using machine learning techniques. In our cursory examination, 60% of sites redirected users to parking web pages that contained only ad content, while in the remaining cases the pages contained a non-trivial number of ads in addition to seemingly legitimate content.

With regards to front-running, while we found no statistically significant evidence of misbehavior by individual registrars during the course of this study, we uncovered the fact that many registrars have several subsidiaries that also perform registration duties on their behalf. The connections among these entities are exceedingly difficult to discover and, unfortunately, little information exists in the public domain that can be used to confirm them. Therefore, if some registrars were involved in front-running behaviors, it is entirely possible that they could hide questionable activities by routing registrations through subsidiaries or partners. As a whole, these results call into question overhasty statements by ICANN that front running is not occurring. Moreover, from what we can tell, these relationships frequently change, underscoring the difficulty in detecting misbehavior by dishonest registrars. Overall, our findings seem to indicate a need for policy-based responses to the problem, rather than technological solutions.

Regarding Mitigation. Clearly, there are two potential options for mitigating the types of domain registration abuses described in this paper: (1) technological methods for identifying abusive web pages, and (2) changes to ICANN policy that disincentivize the behavior. Regarding technological mitigation strategies, Moore *et al.* [14] illustrate that it is indeed possible to identify ad-centric web pages using their content. By removing these pages from search engines, as Google did in 2011, it is possible to appreciably drop their ad revenue, thereby making them unprofitable for the domain owner. However, if we consider the lessons learned from the struggle to stop spam, then it is clear that it is only a matter of time before these ad-centric web pages adapt to the detection mechanisms and make their content even harder to distinguish from legitimate web pages (an already difficult task, as discussed above). Moreover, this type of detection does little to stem the revenue derived from the opportunistic resale of these domains to their respective copyright owners or legitimate content producers.

Policy-oriented mitigation, on the other hand, presents an interesting opportunity for preventing this behavior due primarily to the centralized nature of the domain name registration system. Unfortunately, while it is clear that policy changes instituted by ICANN had an appreciable impact on the practice of domain tasting, reforms aimed at curtailing speculation and front-running appear to be non-existent. One obvious, if drastic, solution might be to eliminate the conflict of interest that arises when registrars are allowed to sell domain names. Other potentially effective approaches, including offline domain availability

checks, have also been put forth. However, these approaches have all been rejected outright by ICANN, even after seemingly acknowledging the threat of domain speculation as the reason for postponing any new applications for generic top-level domains (gTLDs) [19]. A similar concern about domain speculation appears to have motivated the pre-registration period of .XXX domain names [18].

As our results and those of Liu *et al.* [13] indicate, systemic intervention by ICANN appears to be the only reasonable solution to these problems since it is exceedingly easy for those abusing the domain name system to adjust their behaviors in response to policy changes from independent registrars. At the very least, we hope that our results shed light on the challenges inherent in detecting such malfeasance, and that they will spur constructive dialog on relevant public policy.

Acknowledgements. This work was supported in part by the U.S. Department of Homeland Security under Contract No. FA8750-08-2-0147, and the National Science Foundation under award numbers 0831245 and 0937060. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect those of the NSF.

References

[1] D. Kesmodel, The Domain Game: How People Get Rich From Internet Domain Names, Xlibris Corporation, 2008.

[2] Y. Wang, D. Beck, J. Wang, C. Verbowski, B. Daniels, Strider Typo-Patrol: Discovery and Analysis of Systematic Typo-Squatting, in: Proceedings of USENIX SRUTI, 2006, pp. 31–36.

[3] A. Banerjee, D. Barman, M. Faloutsos, L. N. Bhuyan, Cyber-Fraud is One Typo Away, in: Proceedings of the 27^{th} Conference on Computer Communications, 2008, pp. 1939–1947.

[4] T. Moore, B. Edelman, Measuring the Perpetrators and Funders of Typosquatting, in: Proceedings of Financial Cryptography and Data Security, 2010, pp. 175–191.

[5] S. Finkelstein, Domains with Typographical Errors – A Simple Search Strategy, `http://sethf.com/domains/typos/` (January 2003).

[6] S. Finkelstein, Domains with Typographical Errors – A Google Search Strategy, `http://sethf.com/domains/typos-google/` (February 2003).

[7] E. Gabrilovich, A. Gontmakher, The Homograph Attack, Communications of the ACM 45 (2) (2002) 128.

[8] T. Holgers, D. E. Watson, S. D. Gribble, Cutting Through the Confusion: A Measurement Study of Homograph Attacks, in: Proceedings of the 24^{th} Annual USENIX Technical Conference, 2006, pp. 261–266.

[9] Internet Corporation for Assigned Names and Numbers, The End of Domain Tasting: Status Report on AGP Measures, `http://www.icann.org/en/tlds/agp-status-report-12aug09-en.htm` (August 2009).

[10] Internet Corporation for Assigned Names and Numbers, Report on Domain Name Front Running, `http://www.icann.org/en/committees/security/sac024.pdf` (February 2008).

[11] B. Edelman, Front-Running Study: Testing Report, `http://www.icann.org/en/compliance/edelman-frontrunning-study-16jun09-en.pdf` (June 2009).

[12] LegitScript and KnujOn, Rogues and Registrars: Are Some Domain Name Registrars Safe Havens for Internet Drug Rings?, `http://www.legitscript.com/download/Rogues-and-Registrars-Report.pdf` (April 2010).

[13] H. Liu, K. Levchenko, M. Felegyhazi, C. Kreibich, G. Maier, G. M. Voelker, S. Savage, On the Effects of Registrar-Level Intervention, in: Proceedings of the 4^{th} USENIX Workshop on Large-Scale Exploits and Emergent Threats, 2011.

[14] T. Moore, N. Leontiadis, N. Christin, Fashion crimes: Trending-term exploitation on the web, in: Proceedings of the 18^{th} Annual ACM Conference on Computer and Communications Security, 2011.

[15] VeriSign, Inc., VeriSign TLD Zone Access Program, `http://www.verisign.com/domain-name-services/domain-information-center/tld-zone-access/` (2009).

[16] S. E. Coull, A. M. White, T.-F. Yen, F. Monrose, M. K. Reiter, Understanding Domain Registration Abuses, in: International Information Security Conference, 2010, pp. 68–79.

[17] The domain name industry brief, VeriSign, Inc. (2011).

[18] D. Coldewey, Lock Down Your .XXX Domain Before The Land Rush Begins, `http://techcrunch.com/2011/09/07/lock-down-your-xxx-domain-before-the-land-rush-begins/` (September 2011).

[19] M. Palage, New gTLDs: Let the Gaming Begin, Progress on Point 16 (2009) 1–9.

[20] R. Agrawal, T. Imieliński, A. Swami, Mining Association Rules Between Sets of Items in Large Databases, in: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 1993, pp. 207–216.

[21] D. Arthur, S. Vassilvitskii, k-Means++: The Advantages of Careful Seeding, in: Proceedings of the 18^{th} Annual ACM-SIAM Symposium on Discrete Algorithms, 2007, pp. 1027–1035.

[22] E. Zivkovic, The Beginner's Domain Name and Cash Parking Guide, `http://formworkblog.com/dl/domain-guide.pdf` (2007).

[23] M. H. Quenouille, Approximate Tests of Correlation in Time-Series, Journal of the Royal Statistical Society. Series B (Methodological) (1949) 68–84.